# Rotation-blended CNNs on a New Open Dataset for Tropical Cyclone Image-to-intensity Regression*

Boyo Chen
Department of CSIE
National Taiwan University
Taipei, Taiwan
r05922050@ntu.edu.tw

Buo-Fu Chen
National Center of Atmospheric Research
Boulder, CO, USA
bfchen751126@gmail.com

Hsuan-Tien Lin
Department of CSIE
National Taiwan University
Taipei, Taiwan
htlin@csie.ntu.edu.tw

## ABSTRACT

Tropical cyclone (TC) is a type of severe weather systems that occur in tropical regions. Accurate estimation of TC intensity is crucial for disaster management. Moreover, the intensity estimation task is the key to understand and forecast the behavior of TCs better. Recently, the task has begun to attract attention from not only meteorologists but also data scientists. Nevertheless, it is hard to stimulate joint research between both types of scholars without a benchmark dataset to work on together. In this work, we release a such a benchmark dataset, which is a new open dataset collected from satellite remote sensing, for the TC-image-to-intensity estimation task. We also propose a novel model to solve this task based on the convolutional neural network (CNN). We discover that the usual CNN, which is mature for object recognition, requires several modifications when being used for the intensity estimation task. Furthermore, we combine the domain knowledge of meteorologists, such as the rotation-invariance of TCs, into our model design to reach better performance. Experimental results on the released benchmark dataset verify that the proposed model is among the most accurate models that can be used for TC intensity estimation, while being relatively more stable across all situations. The results demonstrate the potential of applying data science for meteorology study.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**;

## KEYWORDS

Atmospheric Science, Tropical cyclone, Tropical cyclone intensity, Convolutional Neural Network, Regression, Blending, Dropout, Pooling

*dataset at https://www.csie.ntu.edu.tw/~htlin/program/TCIR


*KDD '18, August 19–23, 2018, London, United Kingdom*

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219926

**ACM Reference Format:**
Boyo Chen, Buo-Fu Chen, and Hsuan-Tien Lin. 2018. Rotation-blended CNNs on a New Open Dataset for Tropical Cyclone Image-to-intensity Regression. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219926

## 1 INTRODUCTIONS

Tropical cyclone (TC, also called as typhoon, hurricane, or cyclone) is a type of severe weather system that forms and develops on the warm tropical ocean. If a TC makes landfall, it could cause a significant threat to life and property. The intensity of a TC, which is defined as the maximum sustained surface wind near the TC center, is one of the most critical parameters for TC warning and disaster management. For instance, [30] showed that the power relationship between TC intensity and damage might range between 4 and 12. Therefore, pursuing an accurate estimation of TC intensity is an important task for meteorologists and weather forecasters.

Note that in-situ measurement of TC intensity is difficult because TCs spend most of their lifetime on the open ocean. Therefore, the observations from satellite remote sensing serve as primary sources of TC information due to their global coverage and high temporal frequency. Although satellite remote sensing is not capable of directly measuring the wind near the surface, satellite imagery of cloud, water vapor, and precipitation can be used as proxies for estimating TC intensity indirectly [19, 27].

In meteorology, the development of TC intensity estimation nowadays highly relies on constructing informative features for the estimation task. In particular, Section 2 introduces several models based on first calculating informative features and parameters from the satellite images, and then apply various types of regression models to solve the estimation task. Nevertheless, even for experienced meteorologists, it can be difficult to identify informative features for the diverse TCs in different life stages and in different basins. Many current models thus rely on relatively few human-constructed features (usually fewer than 10), which makes the models somewhat restricted in capacity.

In the data science side, ever since AlexNet was proposed in 2012 [11], deep learning techniques have been flourishing for various types of estimation tasks. One important advantage of deep learning techniques is automatic feature construction, which has started to successfully replace human-constructed

features in many estimation tasks, such as object recognition. In this study, we aim to replicate the success by applying Convolutional Neural Network (CNN) in deep learning to solve the TC-image-to-intensity regression task (shorthanded "image regression" task). The study provides an alternative to current models that rely on human-constructed features.

Recently, tasks related to our image regression task has begun to draw attentions from data scientists [15, 21]. To facilitate the data scientists in studying the image regression task together with meteorologists, we also collect and publish an open benchmark dataset in this study. The details of the dataset will be described in Section 3. Then, Section 4 describes our proposed CNN model for the image regression task. The model is extensively compared against state-of-the-art models for the image regression task in Section 5. The results demonstrate that the proposed model outperforms most of the state-of-the-art models, and justify the validity and potential of applying CNN on the image regression task. We summarize our findings in Section 6.

## 2 RELATED WORK

One of the basic ideas of using satellite imagery to estimate TC intensity is that TCs associated with similar cloud features may have similar intensity. The Dvorak technique is the most widely used methodology that estimates TC intensity based on TC cloud features observed from geostationary satellites [2, 27]. This methodology correlates TC intensities to various cloud patterns of central and banding features in the infrared images. However, it takes significant time to master in the Dvorak technique and its regional nuances and adjustments. Furthermore, the inherent subjectivity of the storm center selection and scene type determination procedures compromises the stability of the forecasting skill. During the last three decades, several revised Dvorak techniques have been developed [18, 29]. The Advanced Dvorak Technique (ADT) [18, 19] is the latest released version and currently used for operational TC intensity estimation. ADT reduces the subjectivity by using computer-based algorithms for recognizing cloud features. Besides, ADT is about the first one to apply linear regression on estimating TC intensity.

In additional to Dvorak-based techniques, other parameters calculated from the infrared satellite images are proposed to correlate with TC intensity. A promising and relatively new parameter is the deviation angle variance (DAV), which determines the symmetry level of a TC by evaluating the gradient of the cloud top temperature [22, 23]. Besides, several other parameters are also used for correlating TC cloud feature to TC intensity, such as mean and standard deviation of cloud top temperature in 14 radius ring around a TC [3], and slope of TC cloud top in the inner-core [24].

Meanwhile, ever since ADT [18] took linear regression into the model, various regression methods have also been applied for TC intensity estimation. Here are some examples:

(1) [22] used a nonlinear sigmoid equation to describe the relationship between DAV and TC intensity.
(2) [3] used the k-nearest-neighbor algorithm on TCs with similar intensity based on satellite images.
(3) [32] introduced a multiple linear regression model that considers seven different parameters of TC cloud characteristics.
(4) [31] introduced a machine learning method called relevance vector machine (RVM) for TC intensity estimation.

Recently, some researchers started to apply CNN in classifying TCs by intensity. In 2017, [15] divided TCs into classes based on maximum sustained wind speed with 5-knots intervals. They first pre-trained their model on ImageNet, then fine-tune the model by the cross-entropy loss on the TC classification dataset. In 2018, [21] divided TCs into 8 categories, each with physical meanings in meteorology. The intervals between these categories are ranged from 12 to 29 knots. The authors also visualized the features extracted from CNN. Both works have reached remarkable results. Nevertheless, arguably solving the image regression task is different from solving a classification task. First, compared with the regression task, a classification task loses some information because different intensities are clipped into one interval. Second, the classification task generally does not take the magnitude of the "classification" error into account. That is, misclassifying a 80-knot TC to a 30-knot-class is the same as misclassifying the TC to a 70-knot-class. The difference calls for a true end-to-end regression model for the TC intensity estimation task.

## 3 THE TROPICAL CYCLONE FOR IMAGE-TO-INTENSITY REGRESSION DATASET (TCIR)

Although atmospheric researchers started to estimate TC intensity using satellite data since the 1970s, TC data still remains less accessible to most data scientists. To let people start investigating TC estimation tasks more conveniently, we collect the Tropical Cyclone for Image-to-intensity Regression (TCIR) dataset. TCIR serves as an open benchmark dataset for data scientists to evaluate TC intensity estimation models fairly. The dataset can be downloaded at

https://www.csie.ntu.edu.tw/~htlin/program/TCIR

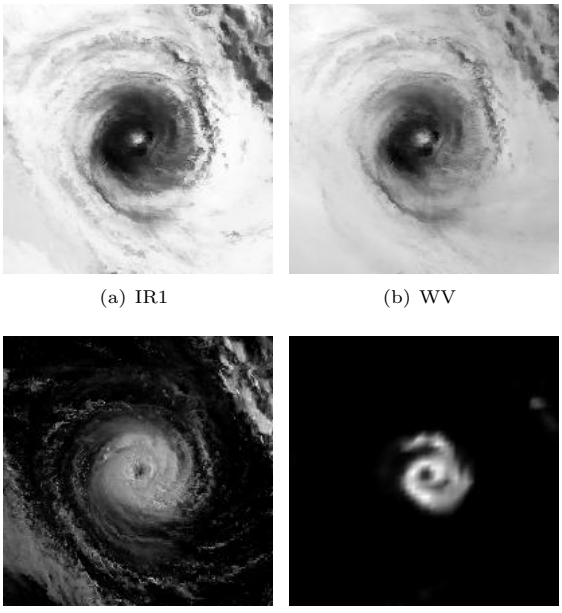
along with simple usage explanations.

### 3.1 Sources

Satellite observations comprising TCIR are from two open sources:

**GridSat** [4, 8], *Gridded Satellite Data*: GridSat is a long-term dataset of global infrared window brightness temperatures, including three channels: IR1, WV, and VIS. This dataset includes data from most meteorological geostationary satellites every three hours since 1981. The resolution is 7/100 degree latitude/longitude.

**CMORPH** [6], *CPC MORPHing technique*: CMORPH precipitation rates from 2003 to 2016 were included into TCIR. CMORPH provides global precipitation analyses at relatively



(a) IR1 (b) WV

(c) VIS (d) PMW

**Figure 1: An example of the four channels from TCIR. Note that they are scaled to the range [0, 256) and drawn in gray scale.**

high spatial and temporal resolution, which uses precipitation estimates derived from low orbit microwave satellite observations exclusively and whose features are transported via spatial propagation information obtained entirely from geostationary satellite IR1 data. The resolution of CMORPH is 0.25-degree every three hours.

### 3.2 Channels

Note that because CMORPH was unavailable before 2002, TCIR includes data starting from 2003. Each data point contains 4 channels, including

0 IR1: Infrared
1 WV: Water vapor
2 VIS: Visible light channel
3 PMW: Passive micro-wave rainrate

In Fig. 1, we scale each channel to the range [0, 256) and draw them as gray scale images. The images allow us to understand each channel better. From the figure, we can see that IR1 and WV provide similar information. Also, a closer look reveals that the VIS channel is very unstable because of the daylight situation. In particular, more than half of the frames during the night have very noisy VIS channel images.

### 3.3 Frames

A total of 47916 frames from 861 TCs in western North Pacific, eastern North Pacific, and Atlantic ocean are collected by TCIR (Table 1).[1]

[1]After the acceptance of KDD, we have added TCs in some other regions, such as the Southern Hemisphere.

| Region | # TCs | # Frames |
|---|---|---|
| West Pacific | 379 | 20060 |
| East Pacific | 247 | 14149 |
| Atlantic | 235 | 13707 |
| Total | 861 | 47916 |

**Table 1: Data Amount included in TCIR**

For each frame, there are $201 \times 201$ data points, and the resolution is 7/100 degree lat/lon. That is, the width and height of a frame are 14 degrees lat/lon, and the distance between 2 grids is about 4 km. Note that the center of TCs are all placed at the middle grid of each frame. Also, the original resolution of the PMW channel from CMORPH is 1/4 degree lat/lon, to unify the size of all 4 channels, we rescale PMW channel to be about 4 times larger by linear interpolation.

Furthermore, because the data are collected by satellite, some of the values could be missing or damaged. We just left them there as NaN. It is up to the data scientists to decide how to handle those NaN value when feeding the dataset into models.

### 3.4 Other TC information

TCIR also provides the information from TC best-track data, which could be considered as the truth when developing the regression models. We used the best-tracks from Joint Typhoon Warning Center (JTWC) for TCs in western North Pacific (WP); and the best-tracks from the revised Atlantic hurricane database (HURDAT2) for TCs in eastern North Pacific (EP) and Atlantic Ocean (AL) from 2003 to 2016. The TC information provided in TCIR includes TC center location, intensity, (i.e., the maximum sustained wind, in kt), minimum sea-level pressure, and size (i.e., the mean of radii of 35-knot wind in the four quadrants, in nmi). Note that these values are tuned and finalized afterward based on all observation that is available, and can be very different from the real-time estimations in meaning. We believe that estimating the size can be the next important task for estimating TC property from images.

While the best-track information can be taken as ground truth, they are still some "estimation" in nature and can suffer from some inherent noise. Thus, for the image regression task, an error within 10 knots is generally acceptable.

## 4 PROPOSED METHOD

We start by pointing out the properties of our image regression task. In particular, we discuss its differences to the image classification task that comes with many mature techniques, and we discuss its known specialties obtained from the meteorologists. We then illustrate how the properties can be leveraged to properly design our proposed model.

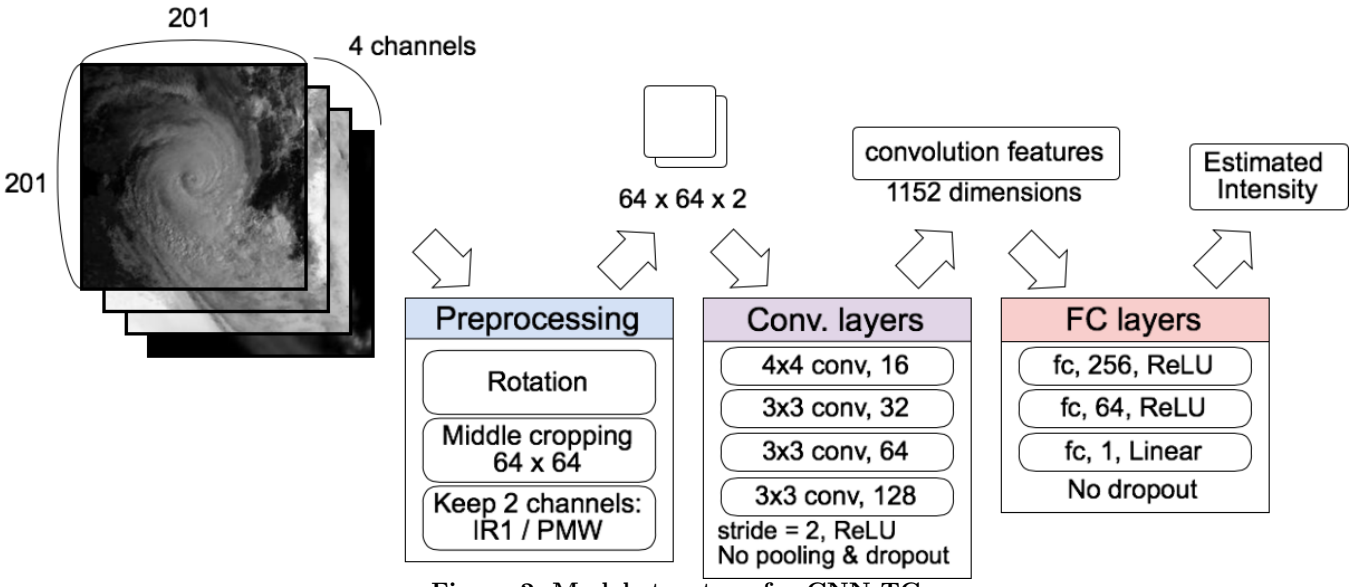



**Figure 2: Model structure for CNN-TC**

### 4.1 Comparison to image classification

Convolutional Neural Network (CNN) is a mature family of deep learning models for image classification tasks such as digit identification [12] and object recognition [10]. This work extends from those mature models to solve our image regression task. In our initial study, we directly applied some popular and classical CNN models such as AlexNet [11] and VGGNet [25] on our image regression task. We then observed some major differences between our task and typical image classification ones, and proposed some careful modifications to address the differences and improve the performance. We summarize those differences in four aspects below.

**Output.** The most obvious difference between our task and the classification one is the type of output variable. Classification aims for a discrete output within a finite number of classes. The goal is generally transformed to producing a soft vector that represents the conditional probability of each class given the image, and the maximum-probability class can be taken as the output. The cross-entropy loss function is commonly used to measure the difference between the predicted soft vector and the one-hot-encoded class identity. [11, 25]

Because only the maximum within the soft vector is taken as the discrete output, it suffices to focus on the *relative* magnitude within the soft vector for accurate classification. Our task, however, demands predicting the *actual* magnitude of the intensity, which is a continuous variable. We take the mean squared error (MSE) as the loss function in our task, and will discuss how the change of loss function affects our model design in Section 4.2 and Section 4.3.

**Target.** Many image classification tasks aim at locating particular object(s) within the image within some particular area. The objects can often be distinguished from the background by clear contours. For our image regression task, the TC images are mashy, continuous and do not come with contours. Then, techniques like (maximum) pooling in image classification CNNs cannot be used to enhance the activation of the object contour over the background. We will discuss this issue in more detail in Section 4.3.

**Position.** Image classification tasks are usually less sensitive to the location of the objects within the image. Thus, it is possible to preprocess the original dataset by augmenting randomly-cropped version of the images to improve the performance of the classification model, as done by AlexNet [11].

Our image regression task, however, comes with an aligned dataset where the cyclone center is within the middle of every TC image. According to the domain knowledge of meteorologists, critical factors for predicting the intensity usually lies within a range of 2 degrees (lat/lon) from the center. Random-cropping, which causes the center to shift within the image, will then make it harder for the model to capture the critical factors and can be harmful instead of helpful to the performance. We will discuss other preprocessing steps that help the model capture the critical factors more easily in Section 4.4.

**Orientation.** In addition to random cropping, other data augmentation techniques for image classification include horizontal flipping and small-angle rotations. For instance, VGGNet [25] studied the importance of horizontal flipping. Big-angle rotations are seldom used for image classification because there are common poses for many discrete objects—for instance, a car is usually not upside down.

The TC images, however, should perhaps not be flipped horizontally, because flipping violates an important domain knowledge that all TCs in the Northern Hemisphere should rotate counterclockwise, given that the TCs used in this study are all in the Northern Hemisphere. On the other hand, it appears possible to rotate the TC images by arbitrary degrees

to hint the "symmetry" of different parts of the image for contributing to the intensity prediction. The idea of using rotation for augmentation leads to a significant improvement in performance, and will be discussed further in Section 4.4.

Considering the quality and quantity of the images within the TCIR dataset, we decide to take the AlexNet [11] as our base CNN structure for this initial study instead of deeper ones like VGGNet [25] or GoogLeNet [26]. We then modify AlexNet to our proposed CNN-TC model by addressing the four aspects of differences above. Meanwhile, we also lighten the structure so that no pre-training is needed. In the next sections, we illustrate our modifications, as depicted in Fig. 2.

## 4.2 Initialization

Properly initializing the weights of deep models like CNN allows the models to start learning in a better mode without being trapped at bad local optima or flatlands. A common and simple strategy is to initialize each weight randomly from a normal distribution with zero mean and a standard deviation within $[0.1, 0.5]$. Recently, it is shown that a standard deviation of $\sqrt{2/\ln d}$, where $d$ is the number of inputs, leads to robust performance [20]. There are also more sophisticated strategies like layer-sequential unit-variance (LSUV) [16] for assisting the deep models.

The strategies discussed above have been studied mostly for image classification tasks, which only care about the relative magnitude between classes within the soft vector, as discussed. For our image regression task that requires the actual magnitude to be accurate, it turns out that the strategies above make the weights and the outputs too large, causing slow and unstable convergence.

After some studies of different initialization strategies, we figure out that the randomly-normal strategy with a standard deviation of 0.01, which is much smaller than those being used for image classification, suffices to ensure stable convergence. For simplicity, we take the 0.01-standard-deviation strategy for our CNN-TC model in all experiments.

## 4.3 Removal of Pooling and Dropout

**Removal of pooling.** The pooling technique in standard CNNs aims to reduce computational complexity by keeping the statistics of a group of features instead of their original values. There are two popular pooling strategies: maximum-pooling and average-pooling. The former keeps the maximum within the group, which effectively emphasizes more important information like contours. The latter keeps the average of the group, which removes variations within the group to possibly be more noise-resistant. We have discussed in Section 4.1 that maximum-pooling may not be effective for our image regression task because the lack of contours. Next, we illustrate the situation in more detail, and provide toy examples to explain why the average pooling may also not be effective.

Note that every patch within the TC image could provide information on the intensity of the TC. For instance, a patch that contains no clouds is as important as a patch that is full of steady clouds in intensity prediction. Consider the three matrices below with values representing the density level of clouds at some locations:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 4 & 1 \\ 1 & 4 & 4 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 4 & 4 & 0 \\ 0 & 4 & 4 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Arguably the first two "TCs" are different because the former comes with more clouds in total. But if we apply a $2 \times 2$ maximum-pooling on the matrices, the first two matrices become indistinguishable, as follows. The toy example demonstrates the potential harmfulness of applying maximum-pooling in terms of dropping some secondary "magnitude" information.

$$\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}, \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

On the other hand, the last two "TCs" are also different because one comes with a larger cloud density near the center. If we apply a $2 \times 2$ average-pooling on the matrices, the last two matrices become indistinguishable, as follows. The toy example demonstrates the potential harmfulness of applying average-pooling in terms of dropping the "contrast" information.

$$\begin{bmatrix} 1.75 & 1.75 \\ 1.75 & 1.75 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Because of the potential harmness of the pooling techniques on our image regression task, we decide to remove all pooling layers in our proposed CNN-TC model.

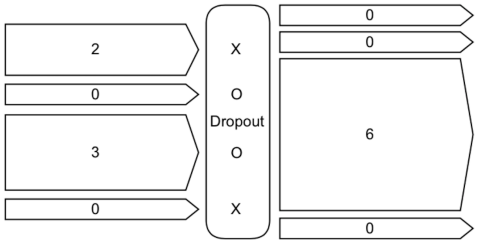



**Figure 3: A simplified illustration of dropout layers**

**Removal of dropout.** Dropout is a mature regularization technique for deep learning. Figure 3 shows a simplified illustration of how dropout works. Suppose that we have a keep rate $p = 50\%$, in the training stage, we will randomly drop a individual nodes in the probability of $1 - p = 0.5$. But as for keeping the expected value for each node to be the same, we will multiply the remain vector by the factor of $p^{-1} = 2$. This operation help us to maintain L1 distance to be about the same. Nevertheless, the L2 distance would become approximately $\sqrt{p}$ times larger.

As mentioned in Section 4.1, we focus on *actual* magnitude of the intensity in this task. But with the usage of dropout layers, regressor would get inputs with longer L2 length during the training stage. Therefore, when we stop to dropout in validation stage, we conjecture that the feature vector became

smaller in L2 length and the outcome estimation would be more likely to under estimate. We will verify the conjecture with some experiments in Section 5.2.

Several works have found that dropout layers can cause some troubles, such as [5] when coupled with batch normalization, and [13] for the trouble of variance shift. Based on the verification of our conjecture above and the troubles in the literature, we decide to remove all dropout layers completely while using the data augmentation techniques described below for regularization.

### 4.4 Preprocessing

Before feeding images into the CNN-TC model for training, we have 3 important preprocessing steps: dropping 2 channels, middle-cropping, and random rotating.

According to Fig. 1, the information provided by IR1 and WV are similar. We thus decide to drop one of them to prevent accidental overfitting on correlated features. Given that IR1 is often a more critical channel for meteorologists to judge intensity, we keep IR1 and drop WV. Meanwhile, since the quality of VIS channel is unstable and strongly affected by the daylight, we decide to drop VIS channel in this study, leaving its potential for future studies.

The reason of middle-cropping is to drop less-important area in the data. For example, in Fig. 1, we can see that this TC only lied in the center part of the frame. From the domain knowledge of meteorologists, critical factors of intensity lie in the area which are less than 2 degrees (lat/lon) from the center. Recall that the distance between 2 nearby points in TCIR is 0.07 degree. $4/0.07 \approx 57.1$. To make sure we discard no important information, we crop the image to $64 \times 64$, which is slight larger than 57. And since TC centers are already placed in the center, we can conveniently crop from the center part of image.

As mentioned in Section 4.1. We need different ways to do data augmentation in this image regression task. Taking advantage of the property that TC data is rotation invariant, we rotate the TC images by arbitrary degrees before training. Rotating the original image would cause white spaces in the corners. Nevertheless, the issue can be easily solved by performing rotation before middle-cropping. The rotation acts as an alternative regularization technique (to the dropout one that we have removed).

### 4.5 Blending by rotation

Given the capability of the proposed CNN-TC model to accept rotated images, we decide to systematically leverage rotation with a mature machine learning technique shows blending. Blending multiple estimations properly is known to reduce the variance of estimation towards more stable performance. Here we simply apply blending the estimations of the models trained with images rotated by different angels. We rotate our images by evenly distributed angles ranged from 0-360. That is, if we plan to blend the estimations from 4 models, each will be trained by images rotated with 0, 90, 180, 270 degrees, respectively. In Section 5.3, we conduct experiments to test the effect of blending different estimations.

## 5 EXPERIMENT AND ANALYSIS

In this section, we first conduct experiments to show how we can get better results by removing dropout layer and use rotating instead to fight against overfitting. Second, we explore the effect of rotation-blending.

Then, we compare CNN-TC with some models which are now used in operational applications. Our results reveal a high potential for CNN-TC to be used in real-world applications. Also, we compared CNN-TC with other works for intensity estimation. The result also shows that our model can reach better performance.

### 5.1 Experiment Settings

Before training, to let the model learn more easily, we standardize every value by

$$x_{\text{standardized}} = \frac{x - \text{mean of the channel}}{\text{standard deviation of the channel}}$$

As TCIR is originally collected from satellite, there exist some damaged values. There are two groups of them, one is NaN values, and another is extremely large values. Since we standardized data with mean and standard deviation, to prevent mean from being strongly impacted by these large values, we replace a value with 0 if it is originally greater than $10^3$. For those NaN values, we directly assign them as 0.

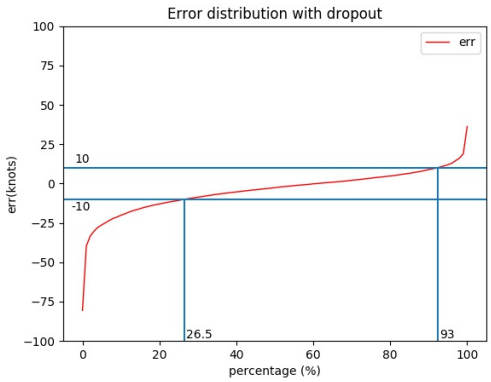
We use the data of TCs during 2003-2014 as training data and TCs during 2015-2016 for validation. There are 40348 frames from 730 TCs in the training data and 7569 frames from 131 TCs in the validation data. We take the following key parameters:

(1) Epochs numbers: stop at 500 epochs
(2) Learning rate: $5 \times 10^{-4}$
(3) Regularization parameter: $10^{-5}$ for every weight.

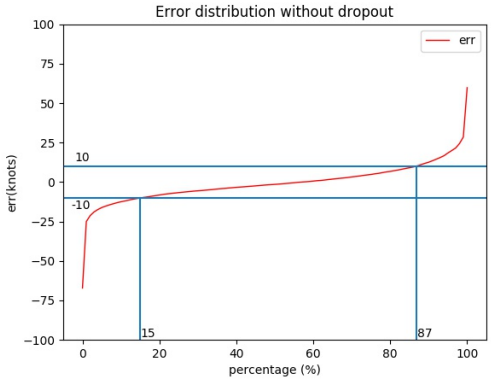
### 5.2 Dropout

As mentioned in Section 4.3, we think that a dropout layer could influence the feature's L2 distance during training stage and thus make the model more likely to underestimate during validation stage. In this experiment, two models were trained in identical structures and epoch numbers, except that one with dropout and another without dropout. We calculated error distance from 7569 validation frames and sort them in ascending order, so that we can know the error distance at every percentage. The result is shown in Fig. 4.

We can clearly observe that when training with dropout layers, 26.5% frames are underestimated by more than 10 knots, while only 7% frames are overestimated by more than 10 knots. But after removing dropout layers, the error distributed like a Gaussian distribution, divided into both sides evenly. We also discover that with more epoch trained, the level of underestimate became more serious. Therefore, we removed dropout layer from CNN-TC.




(a) With dropout




(b) Without dropout

**Figure 4: Line chart of error distribution. In the chart, we can observe that using dropout in regression task would cause underestimate.**

## 5.3 Rotate and Blending

Rotation-blending is a special but intuitive design which can hardly apply to other classical image classification task. To show the effect of rotation-blending, in this experiment, we test different numbers of estimations to be blended. The result is shown in Fig. 5.

Comparing to the learning curve of not using rotation-blending, we can see that rotation-blending provided remarkable improvements. Moreover, with more estimations we used for blending, we can receive better results. The improvement slows down after we extend our blending number to over 9. To balance between computation loading and performance, we use 10 as our blending number.

## 5.4 Results

In this subsection, we will use TCs during 2015-2016 to evaluate our performance.

First, our result is compared to three models, which are used in operational intensity estimations nowadays: ADT, AMSU, and SATCON. We collected their raw estimations

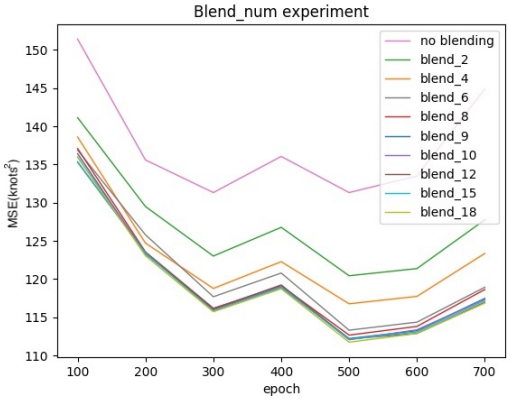



**Figure 5: Learning curve of validation MSE using different number of blending estimations. Notice that we optimized our model only on a single rotated image, blending is completed during the validation stage.**

from Cooperative Institute for Meteorological Satellite Studies websites (CIMSS) [1].

In Fig. 6, X-axis represents the true values and Y-axis represents the estimated values. Notice that the frame count $n$ is different for 4 methods, the reason will be described below.

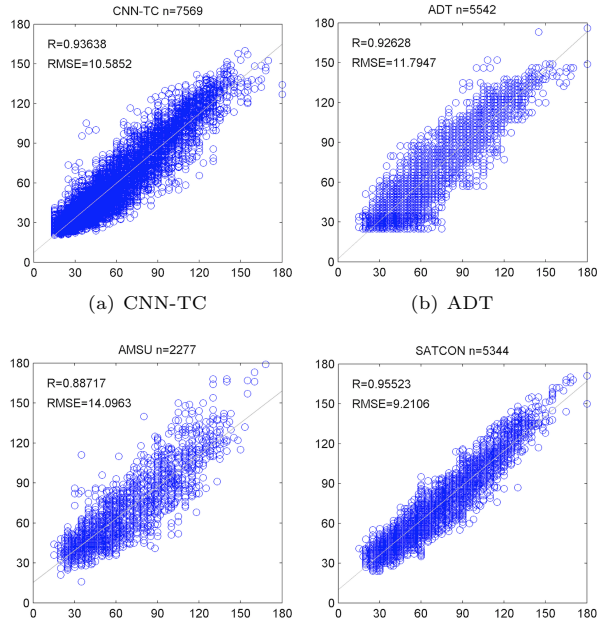
ADT [17, 18] (Advanced Dvorak Technique) is generally better for strong TCs, but as for fitting better in stronger frames, it usually overestimates those weaker TCs. Furthermore, when the outcome predictions are less than a threshold (about 25 – 30 knots), ADT will directly abandon those predictions. These traits can be observed clearly in Fig. 6-(b)

AMSU [7] (Advanced Microwave Sounding Unit) uses low earth orbit satellites instead of geosynchronous satellites. Thus, the estimation can be provided only when the satellites ran across the TCs. Comparing to ADT, AMSU is known to be better for weaker TCs than for those stronger ones.

SATCON [28] (SATellite CONsensus) is the combination of ADT, AMSU, and some other minor methods. It could be taken as some heuristic blending guidelines depending on ocean region, estimated intensity level from each model, and more. SATCON is a model that highly rely on expert's human intelligence, and is unable to estimate TC intensity when estimation from ADT is unavailable or AMSU samples are not enough.

In Table 2, we show the average RMSE (root mean square error) for each model and each region. Notice that ADT and SATCON both have fine-tuned their model for each region. Comparatively speaking, CNN-TC provided stable estimation for all cases by a single model. In the future, we can also consider fine-tuning CNN-TC for each different ocean regions to shoot for even better results.

In Fig. 7, we show several examples for full life cycle TCs. Recall that AMSU uses low-Earth-orbit satellites so the estimation is not continuous. CNN-TC is designed as single




(a) CNN-TC (b) ADT

(c) AMSU (d) SATCON

**Figure 6: Comparison with three models for operational intensity estimations. X-axis represents the true values and Y-axis represents the estimated values. Data during 2015 2016 are used for verification**

| | West Pacific | East Pacific | Atlantic | Overall |
|---|---|---|---|---|
| CNN-TC | 12.25 | 9.42 | 9.27 | 10.59 |
| ADT | 12.25 | 10.57 | 12.87 | 11.79 |
| AMSU | 15.68 | 13.76 | 11.39 | 14.10 |
| SATCON | 9.90 | 8.80 | 8.42 | 9.21 |

**Table 2: RMSE for data during 2015-2016 of ADT, AMSU, SATCON, and CNN-TC. ADT and SATCON fine-tuned their model for each region, that is, their result is from 3 different models. In contrast, CNN-TC is a single model optimized for all data from 3 regions.**

frame estimation task. While both of ADT and SATCON have been smoothed, we have not done any smoothing in CNN-TC. However, we can clearly see that after smoothing, CNN-TC can produce an even better performance.

In Fig. 7-(f), the case we show is the second strongest TC in the history. We can observe that ADT perform better when the TC is extremely strong, and SATCON can also rely on ADT to perform well in such condition by heuristic rules. Comparatively speaking, CNN-TC strongly depends on the data amount for training neural networks. When such extremely strong frames are very rare, unavoidable CNN-TC would be weaker in this condition. This issue can possibly be alleviated in the future by using cost-sensitive losses.

| | RMSE (knots) |
|---|---|
| Kossin et al. (2007) | 13.20 [9] |
| FASI | 12.70 [3] |
| Improved DAV-T | 12.70 [23] |
| TI index | 9.34 [2][14] |
| Y. Zhao et al. (2016) | 12.01 [32] |
| J. Miller et al. (2017) | 10.00 [3] |
| R. Pradhan et al. (2018) | 10.18 [4] |
| ADT | 11.79 |
| AMSU | 14.10 |
| SATCON | 9.21 |
| CNN-TC | 10.59 |
| CNN-TC(with smoothing) | 9.45 |

**Table 3: A rough comparison between RMSEs of the models estimating typhoon intensities in other works.**

In Fig. 7-(d)(e)(f), we can observe that ADT and SATCON will give up to estimate sometimes. In contrast, CNN-TC can constantly produce close estimations.

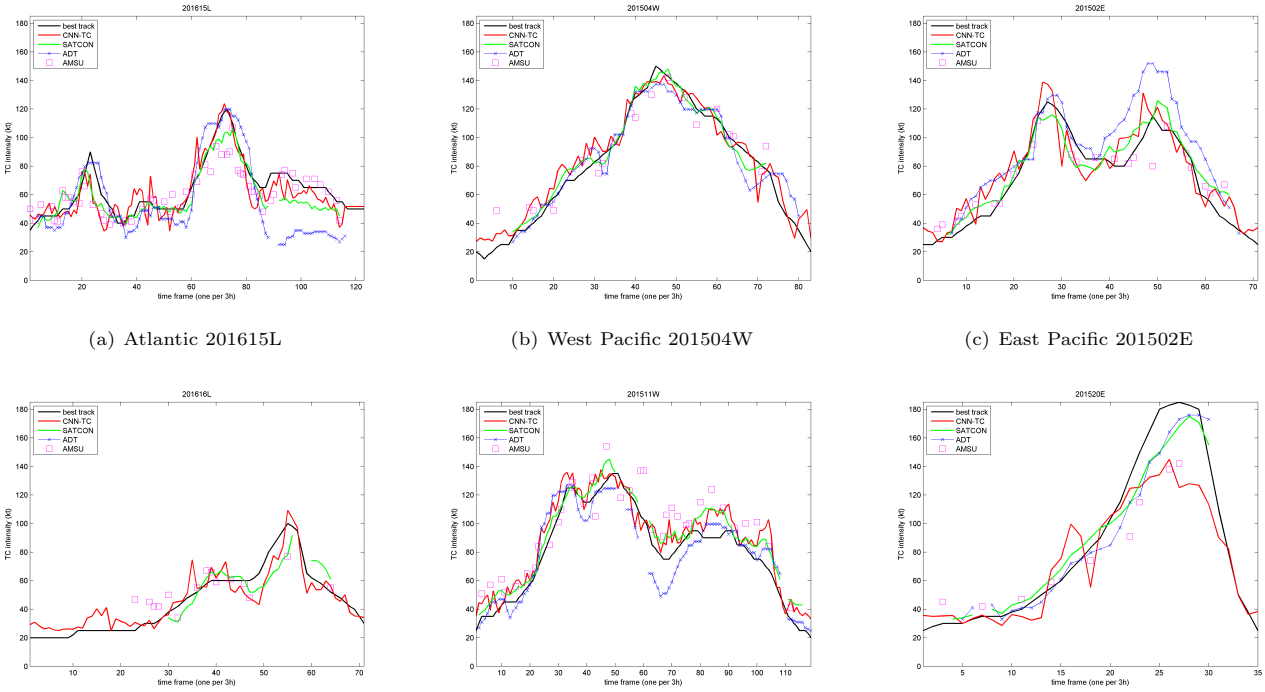
## 5.5 Further enhancement

We designed the CNN-TC model on the image-to-intensity regression task, using only a single frame to estimate the TC intensity. To further test the potential of CNN-TC to be used in real-world applications, here we experiment with a one-sided smoothing, averaging a prediction with 2 previous estimations, and received a significant improvement on our estimation. The result can be seen in Fig. 8.

While one-side smoothing is only a casual method to do smoothing, this experiment suggests that our estimation can be significantly improved even with such an easy smoothing. This also shows the potential for us to extend this task into a sequence learning task for a better result on TC intensity estimation and even forecast.

In the end of this section, we compare our RMSE to that of other works. The details can be found in Table 3. We should keep in mind that these works are evaluated with TC data from different oceans and different years. For example, although TI index exhibits an impressive result, they only use 5 TCs in 2011 for validation. Most model's performances are ranged from 10.00 $\sim$ 13.2 knots in RMSE, while SATCON being the best with 9.21 knots in RMSE. With a simple smoothing, CNN-TC can reach 9.45 knots in RMSE, which is about as good as SATCON. Recall that SATCON depends on other models to make estimations and can thus be unavailable some time, while CNN-TC can constantly produce close estimation by itself. The close RMSE between CNN-TC and

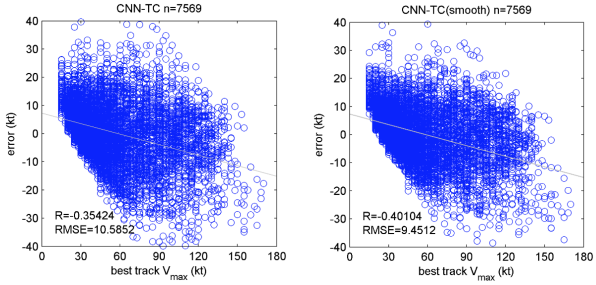
[2]Only 5 TCs in 2011 are used for validation.
[3]Originally classification task, not sure how the estimated intensities are calculated from the classification results.
[4]Originally classification task, the estimated intensities are determined as the weighted average of two highest categories with respect to their probabilities.[21]




(a) Atlantic 201615L (b) West Pacific 201504W (c) East Pacific 201502E

(d) Atlantic 201616L (e) West Pacific 201511W (f) East Pacific 201520E

**Figure 7: Case studies from 6 different TCs. For each ocean regions, we choose 2 TCs to illustrate the robustness of CNN-TC. The black line "best track" stands for "best tracking intensity", are just the label we provided in TCIR. It is revealed that CNN-TC can be further improved by smoothing techniques.**




(a) Unsmoothed (b) Smoothed

**Figure 8: After one-side smoothing with window size=3, we can significantly improve our result from 10.59 to 9.45. The improved result is almost as good as SATCON.**

SATCON and the disadvantages of SATCON makes CNN-TC a preferable model for the image regression task.

## 6 CONCLUSION

In this paper, we dived into the issue of TC intensity estimation. First, we organized a new open dataset TCIR for data scientists to further explore this issue. In this dataset, we collected several remote sensing data from the satellites and organized them into a format with which data scientists would be relatively more familiar.

On the other hand, although meteorologists have already achieved satisfactory results in intensity estimation by feature engineering, we demonstrate the potential of deep learning to hand in an even better result by automatic feature extraction. We modified classical CNN structures to better fit in this image-to-intensity regression task, which became our proposed model CNN-TC. To our best knowledge, CNN-TC is the first CNN model which can directly solve the image-to-intensity regression task. Because of the insufficiency of data in quality and quantity, we lighten our CNN structures to prevent overfitting. CNN-TC not only defeated most of the state-of-the-art models, but also justified its practical superiority by being always available with a low estimation error.

For future works, the TCIR dataset can be used to predicting other labels, such as size (also provided in TCIR as label "R35"), which is closely related to the social impacts of TCs. Previous studies suggested that meteorologists are unable to estimate TC size well so far. This is because we still do not have enough understanding for feature engineering of estimating TC size. In view of the success in this work, it can also be promising to use CNN for estimating the TC size. In summary, this work opens a new door for data scientists to

study meteorology problems with deep learning techniques like CNN.

## ACKNOWLEDGMENTS

We thank the anonymous reviewers and the members of NTU CLLab for valuable suggestions. This material is based upon work supported by the Ministry of Science and Technology of Taiwan under 106-2119-M-007-027.

## REFERENCES

[1] [n. d.]. Cooperative Institute for Meteorological Satellite Studies. http://tropic.ssec.wisc.edu/tropic.php. ([n. d.]). [Online; accessed 8-February-2018].
[2] Vernon F Dvorak. 1975. Tropical cyclone intensity analysis and forecasting from satellite imagery. *Monthly Weather Review* 103, 5 (1975), 420–430.
[3] Gholamreza Fetanat, Abdollah Homaifar, and Kenneth R Knapp. 2013. Objective tropical cyclone intensity estimation using analogs of spatial features in satellite data. *Weather and Forecasting* 28, 6 (2013), 1446–1459.
[4] Anand K Inamdar and Kenneth R Knapp. 2015. Intercomparison of independent calibration techniques applied to the visible channel of the ISCCP B1 data. *Journal of Atmospheric and Oceanic Technology* 32, 6 (2015), 1225–1240.
[5] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. 448–456.
[6] Robert J Joyce, John E Janowiak, Phillip A Arkin, and Pingping Xie. 2004. CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology* 5, 3 (2004), 487–503.
[7] Stanley Q Kidder, Mitchell D Goldberg, Raymond M Zehr, Mark DeMaria, James FW Purdom, Christopher S Velden, Norman C Grody, and Sheldon J Kusselson. 2000. Satellite analysis of tropical cyclones using the Advanced Microwave Sounding Unit (AMSU). *Bulletin of the American Meteorological Society* 81, 6 (2000), 1241–1259.
[8] Kenneth R Knapp, Steve Ansari, Caroline L Bain, Mark A Bourassa, Michael J Dickinson, Chris Funk, Chip N Helms, Christopher C Hennon, Christopher D Holmes, George J Huffman, et al. 2011. Globally gridded satellite observations for climate studies. *Bulletin of the American Meteorological Society* 92, 7 (2011), 893–907.
[9] JP Kossin, KR Knapp, DJ Vimont, RJ Murnane, and BA Harper. 2007. A globally consistent reanalysis of hurricane variability and trends. *Geophysical Research Letters* 34, 4 (2007).
[10] Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. (2009).
[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
[12] Yann LeCun. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* (1998).
[13] Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. 2018. Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift. *arXiv preprint arXiv:1801.05134* (2018).
[14] Chung-Chih Liu, Chian-Yi Liu, Tang-Huang Lin, and Liang-De Chen. 2015. A satellite-derived typhoon intensity index using a deviation angle technique. *International Journal of Remote Sensing* 36, 4 (2015), 1216–1234.
[15] JJ Miller, Manil Maskey, and Todd Berendes. 2017. Using Deep Learning for Tropical Cyclone Intensity Estimation. (2017).
[16] Dmytro Mishkin and Jiri Matas. 2015. All you need is a good init. *arXiv preprint arXiv:1511.06422* (2015).
[17] Timothy L Olander and CS Velden. 2012. The current status of the UW-CIMSS Advanced Dvorak Technique (ADT). In *32nd Conf. on Hurricanes and Tropical Meteorology*.
[18] Timothy L Olander and Christopher S Velden. 2007. The advanced Dvorak technique: Continued development of an objective scheme to estimate tropical cyclone intensity using geostationary infrared satellite imagery. *Weather and Forecasting* 22, 2 (2007), 287–298.
[19] Timothy L Olander and Christopher S Velden. 2009. Tropical cyclone convection and intensity analysis using differenced infrared and water vapor imagery. *Weather and Forecasting* 24, 6 (2009), 1558–1572.
[20] Andre Perunicic. 2017. Choosing Weights: Small Changes, Big Differences. https://intoli.com/blog/neural-network-initialization/. (2017). [Online; accessed 25-July-2017].
[21] Ritesh Pradhan, Ramazan S Aygun, Manil Maskey, Rahul Ramachandran, and Daniel J Cecil. 2018. Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network. *IEEE Transactions on Image Processing* 27, 2 (2018), 692–702.
[22] Elizabeth A Ritchie, Genevieve Valliere-Kelley, Miguel F Piñeros, and J Scott Tyo. 2012. Tropical cyclone intensity estimation in the North Atlantic Basin using an improved deviation angle variance technique. *Weather and Forecasting* 27, 5 (2012), 1264–1277.
[23] Elizabeth A Ritchie, Kimberly M Wood, Oscar G Rodríguez-Herrera, Miguel F Piñeros, and J Scott Tyo. 2014. Satellite-derived tropical cyclone intensity in the north pacific ocean using the deviation-angle variance technique. *Weather and Forecasting* 29, 3 (2014), 505–516.
[24] Elizabeth R Sanabia, Bradford S Barrett, and Caitlin M Fine. 2014. Relationships between tropical cyclone intensity and eyewall structure as determined by radial profiles of inner-core infrared brightness temperature. *Monthly Weather Review* 142, 12 (2014), 4581–4599.
[25] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.
[27] Christopher Velden, Bruce Harper, Frank Wells, John L Beven, Ray Zehr, Timothy Olander, Max Mayfield, Charles Chip Guard, Mark Lander, Roger Edson, et al. 2006. The Dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bulletin of the American Meteorological Society* 87, 9 (2006), 1195–1210.
[28] CS Velden and D Herndon. 2014. Update on the SATellite CONsensus (SATCON) algorithm for estimating TC intensity. *Poster session I. San Diego Google Scholar* (2014).
[29] Christopher S Velden, Timothy L Olander, and Raymond M Zehr. 1998. Development of an objective scheme to estimate tropical cyclone intensity from digital geostationary satellite infrared imagery. *Weather and Forecasting* 13, 1 (1998), 172–186.
[30] Alice R Zhai and Jonathan H Jiang. 2014. Dependence of US hurricane economic loss on maximum wind speed and storm size. *Environmental Research Letters* 9, 6 (2014), 064019.
[31] Chang-Jiang Zhang, Jin-Fang Qian, Lei-Ming Ma, and Xiao-Qin Lu. 2016. Tropical Cyclone Intensity Estimation Using RVM and DADI Based on Infrared Brightness Temperature. *Weather and Forecasting* 31, 5 (2016), 1643–1654.
[32] Yong Zhao, Chaofang Zhao, Ruyao Sun, and Zhixiong Wang. 2016. A Multiple Linear Regression Model for Tropical Cyclone Intensity Estimation from Satellite Infrared Images. *Atmosphere* 7, 3 (2016), 40.